DOCUMENT RESUME

ED 469 462 EC 309 233

AUTHOR	Bielinski, John; Minnema, Jane; Thurlow, Martha
TITLE	A Follow-Up Web-Based Survey: Test and Measurement Expert Opinions on the Psychometric Properties of Out-of-Level Tests. Out-of-Level Testing Report.
INSTITUTION	National Center on Educational Outcomes, Minneapolis, MN.; Council of Chief State School Officers, Washington, DC.; National Association of State Directors of Special Education, Alexandria, VA.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
REPORT NO	NCEO-R-7
PUB DATE	2002-07-00
NOTE	23p.
CONTRACT	H324D990058
AVAILABLE FROM	National Center on Educational Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Rd., Minneapolis, MN 55455 ($5). Tel: 612-624-8561; Fax: 612-624-0879; Web site: http://education.umn.edu/NCEO. For full text: http://education.umn.edu/nceo/OnlinePubs/OOLT7.html.
PUB TYPE	Reports - Research (143) -- Tests/Questionnaires (160)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Disabilities; *Educational Assessment; Elementary Secondary Education; *Gifted; Psychologists; *Psychometrics; Surveys; *Test Reliability; *Test Validity
IDENTIFIERS	*Out of Level Testing

ABSTRACT

A Web-based survey of 25 experts in testing theory and large-scale assessment examined the utility of out-of-level testing for making decisions about students and schools. Survey respondents were given a series of scenarios and asked to judge the degree to which out-of-level testing would affect the reliability and validity of test scores within each scenario. Generally, respondents indicated that the error introduced in the vertical scaling process offset any precision that might be gained through out-of-level testing, especially when students were tested more than two levels below grade. For questions on the validity of decisions about school comparisons, adequate yearly progress, earning a diploma, and instruction, respondents' ratings spanned the continuum from those who indicated that out-of-level testing would dramatically reduce the validity of inferences to those who indicated that out-of-level testing would dramatically enhance the validity of inferences. Most respondents indicated that out-of-level testing would have a detrimental effect on validity, regardless of whether the test was part of a multi-level testing system or a criterion-referenced testing system. The only exception was when test scores were used to guide classroom instruction, in which case 47% of respondents indicated enhanced validity. The survey is appended. (DB)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Out-of-Level Testing Report 7

ED 469 462

A Follow-Up Web-Based Survey: Test and Measurement Expert Opinions on the Psychometric Properties of Out-of-Level Tests

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M.L. Thurlow

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

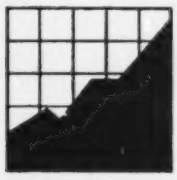
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

EC 309233

2

BEST COPY AVAILABLE

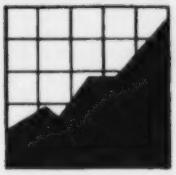
A Follow-Up Web-Based Survey: Test and Measurement Expert Opinions on the Psychometric Properties of Out-of-Level Tests

John Bielinski • Jane Minnema • Martha Thurlow

July 2002

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission, provided the source is cited as:

Bielinski, J., Minnema, J., & Thurlow, M. (2002). *A follow-up web-based survey: Test and measurement expert opinions on the psychometric properties of out-of-level tests* (Out-of-Level Testing Report 7). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Out-of-Level Testing Project is supported by a grant (#H324D990058) from the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NCEO Core Staff

Deb A. Albus
John S. Bielinski
Jane L. Krentz
Kristi K. Liu
Jane E. Minnema
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

Additional copies of this document may be ordered for $5.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://education.umn.edu/NCEO

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

This document is available in alternative formats upon request.

Executive Summary

With the reauthorization of the Individuals with Disabilities Education Act of 1997 (IDEA 97), many states have turned to out-of-level testing as one option for including students with disabilities in statewide testing. The increased dependence on out-of-level testing combined with the high stakes nature of many testing systems has heightened concern about the utility of out-of-level testing for making decisions about students and schools. To address the concerns about the appropriateness of inferences based on out-of-level test scores, we invited researchers, academicians, and others with expertise in testing theory and large-scale assessment to provide their perspectives on this issue. This report summarizes our findings from quantitative analyses of responses to our survey items, and qualitative analysis of the respondents' comments on these issues.

Respondents were given a series of scenarios and asked to make judgments about the degree to which out-of-level testing would have an impact on the reliability and validity of test scores within each scenario. These scenarios dealt with norm-referenced and criterion-referenced tests and included descriptions about linking studies and the specific purposes for which the test scores would be used.

The results show variability in opinions about the impact of out-of-level testing on both reliability and validity. Generally, respondents indicated that the error introduced in the vertical scaling process offset any precision that might be gained through out-of-level testing. The attenuating effect of vertical scaling error on measurement precision was of particular concern when students are tested more than two levels below grade. The opinions of the participants varied even more in the validity section than in the reliability section. For items on the validity of decisions about school comparisons, adequate yearly progress, earning a diploma, and instruction, ratings spanned the continuum from those who indicated that out-of-level testing would dramatically reduce the validity of inferences to those who indicated that out-of-level testing would dramatically enhance the validity of the inferences.

Most respondents indicated that out-of-level testing would have a detrimental effect on validity, regardless of whether the test was part of a multi-level testing system or a criterion-referenced testing system. The only exception was when test scores were used to guide classroom instruction. Yet, just 47% of the respondents indicated that out-of-level testing would enhance validity for instructional decisions.

Many respondents also provided comments. These were examined. Common topics were related to the difficulty of setting an assessment context, the emergence of points of agreement and disagreement, and the role of student-related factors and psychometric test characteristics. Overall, the concerns expressed by respondents and the diversity of opinion among the respondents, highlight the need for caution in using out-of-level testing as well as the need for well-focused research to guide the further development of out-of-level testing policies and practices.

Overview

Out-of-level testing, an approach to measuring student academic skills at the level at which the students are said to be instructed rather than at the grade level in which they are enrolled, was originally used in the 1970s as an indicator of program efficacy for Title I purposes. Since then, out-of-level testing has grown in popularity, although there is little recorded information indicating the extent to which out-of-level tests were administered over the past two decades. Some states (Connecticut, Iowa) have a 10-year or longer history of testing students out of level. At the time we conducted the study reported here, 14 states (Arizona, California, Connecticut, Delaware, Hawaii, Iowa, Louisiana, Mississippi, Oregon, South Carolina, Texas, Utah, Vermont, West Virginia) were testing students out of level as a component of a statewide testing program. The expansion of out-of-level testing, however, did not occur without controversy at the local, state, and federal levels of the educational system. In fact, unresolved issues have prompted one state that had tested students out of level for many years (North Dakota) to recently reverse itspolicy; this state now disallows out-of-level testing in its large-scale assessment programs.

At the time that out-of-level tests were first introduced, testing students out of level was implemented with norm-referenced tests (NRTs). NRTs are developed so that test item content, test item difficulty, and the distribution of test items across content domains are each intended to reflect the curriculum of a representative sample of schools according to specific grade levels. Test companies use a mathematical procedure, called "vertical scaling" to link test items across test (or grade) levels so that the scores of students taking different levels are all reported on a single scale, sometimes called a "developmental scale." Reporting scores on a common scale makes it possible to directly compare the performance of students taking different levels of the test. The developmental scale score is said to have the same meaning regardless of the level of the test a student took. In other words, a score of say 200 is said to reflect the same skills whether a student took the 6th grade test or the 5th grade test. This characteristic of comparability can only be achieved when there is a sufficient degree of overlap of test content between adjacent levels of the test. Generally, there has been little controversy about the utility of developmental scales for making normative comparisons (using NRTs) about students' achievement growth (Minnema, Thurlow, & Bielinski, 2002). This seems to be true even though the amount of error that is introduced through the linking process is unknown (Bielinski, Thurlow, Minnema, & Scott, 2000).

The debate about the appropriateness of out-of-level testing tends to occur when out-of-level test scores are used for either student or system accountability purposes and the tests are not given in adjacent grades so that there is little content overlap across test levels. Today, with the advent of standards-based reform, more states have replaced off-the-shelf tests with criterion-referenced tests (CRTs) developed specifically to measure state content standards, while other

states have augmented their assessment instruments to include both types of measures (NRT/CRTs). These tests are intended to measure proficiency against pre-specified content standards that mark schools' progress toward improving educational results and the student's success in meeting standards. With the requirement that all students be included in large-scale assessment programs, states are pursuing participation options that are less "damaging" to students (Minnema, Thurlow, & Scott, 2001). Out-of-level testing is one mechanism that states are using to increase the participation of students with disabilities in large-scale assessments, believing that measurement of students' skills is also improved (Minnema et al., 2001).

Various groups have expressed concern about out-of-level testing by raising issues about the accuracy (or validity) and precision (or reliability) of out-of-level test results, the difficulty of reporting out-of-level test scores to the public, and the potentially negative effects on the instruction of students who are tested out of level for multiple school years (Minnema et al., 2001; Thurlow, Elliott, & Ysseldyke, 1999). However, because the topic partly is about test theory and scale development, it is important to obtain the perspectives of psychometricians on these issues. The purpose of this study was to obtain input from people with expertise in test theory and scale development about the benefits and drawbacks of out-of-level testing, particularly as they relate to test score reliability and validity.

Method

Participants

The participant pool for this study consisted of persons with expertise in psychometrics who had knowledge of large-scale assessment issues. Many of these individuals, but not all, had participated in focus groups on out-of-level testing. They were identified through nomination from other psychometric experts with whom NCEO staff were acquainted through research publications, or through affiliation with organizations that conduct research on large-scale testing programs. The pool included university faculty, researchers at large educational research organizations, and state and federal education agencies.

Each participant received an email message with an invitation to participate in our study. We received permission to send our survey to 48 individuals. From that group, 25 completed the survey. Some of the other individuals indicated that they did not have sufficient technical expertise on this topic to participate meaningfully in the study.

Instrument

The online survey was designed so that we could obtain meaningful ratings from the participants

of the effects of out-of-level testing on reliability and validity in realistic settings (rather than in abstract, as focus groups did). To assist the participants in making their ratings, three scenarios were provided. The scenarios were based on amalgams of actual statewide testing programs. The scenarios were intentionally brief, providing just what we thought were the critical details that would permit our respondents to make meaningful judgements. Along with each scenario was a set of conditions or constraints intended to provide additional specificity to the scenario.

Participants were asked to provide ratings of the possible effect out-of-level testing would have on test score reliability and validity under each scenario. For the items that addressed the effects on test score reliability, the rating scale had four categories: (1) None (no effect), (2) too little to warrant concern, (3) some, and (4) a lot. For the items that addressed the effects on test score validity, the rating scale had these five categories: (1) dramatically reduce (validity), (2) somewhat reduce, (3) no effect, (4) somewhat enhance, and (5) dramatically enhance. Along with their categorical ratings, narrative boxes were also included for respondents to provide comments that would clarify their ratings if they felt clarification was needed. The full survey is shown in Appendix A.

The online survey was divided into two sections. The first section addressed the degree to which error introduced when linking items across test levels reduces the gain in precision obtained by matching each student to the test that corresponds to their expected performance level. The second section examined the issue of whether out-of-level testing enhanced or degraded the meaning of the scores. Two scenarios were used as a context for ratings.

In the first section (reliability), survey participants were asked to consider the magnitude of the error that is introduced to the item parameter estimates when linking items across different test levels under different types of linking studies, and to indicate whether that error offset any perceived gains in precision achieved through out-of-level testing (see Bielinski et al., 2000).

Participants first considered two types of linking studies. One scenario described a type of linking study typically used by publishers of norm-referenced tests to construct developmental scales. The other was equipercentile method in which a group of examinees takes two levels of the test and the relationship between the scores from the two levels is used to predict the score a student would earn on the in-level test based on performance on the out-of-level test. This second method assumed that the tests were part of a criterion-referenced testing program. Participants were also asked to rate under different levels below grade and whether a locator test or a classroom teacher was used to assign students to levels. A third scenario dealt with the potential biasing effect that emerges because linking studies are conducted primarily on a general education population, whereas those taking out-of-level tests are often only special education students (see Thurlow & Minnema, 2001). Participants were asked to indicate whether this disconnect results in scores on out-of-level tests that are biased.

The second section of the instrument dealt with validity. Two scenarios were used, one based on norm-referenced testing system and one based on criterion-referenced testing system.

Although validity has been characterized many ways, for this study we wanted our experts to provide their insight into whether aggregating out-of-level test scores with in-level scores obscures the meaning of the aggregate. Considering the same scenarios used in Section I, participants were asked to rate whether score interpretation would improve or degrade as the result of out-of-level testing, if the scores were used to: (1) make school-to-school comparisons, or to (2) monitor adequate yearly progress. Several assumptions were made about the testing situation.

For the NRT scenario, it was assumed that test scores were reported on a common scale spanning all levels, that no student was permitted to take a test more than two levels below grade level, no distinction was made in any report between who took which level of the test, and the out-of-level testing rates varied across schools. The issue is whether combining scores for students taking different levels of the test alters the interpretation of the scores.

Respondents were also asked to consider the effect out-of-level testing has on the interpretation of individual scores by classroom teachers. The two situations presented were: (1) using test scores to guide classroom instruction, (2) using test scores to determine whether a student met the passing standard. In the first scenario, it was assumed that the tests were equated and that the scores were reported on a common scale. The issue to ponder here is whether a teacher can make a meaningful evaluation of student performance on the kinds of skills taught in that curriculum given that students took tests with emphasis on different types of skills.

Results

Reliability

The results for the Likert items in the first section are summarized in Table 1. Twenty-two participants responded to each item. There was substantial variability in their opinions on the reliability items. For example, when asked to estimate the degree to which the error in the linking constant offset the gain in measurement precision from out-of-level testing, 9% of the respondents indicated that the linking error would have no effect, and 9% indicated that it would have a large effect. The general consensus was that there was very little to some effect when a student was assigned to a test just one level below grade level. When the test was two levels below grade level, respondents were nearly evenly split between some reduction and a large reduction on the precision of measurement resulting from linking error. For the situation in which students were assigned more than two levels below grade level, the consensus opinion was that linking error had a large effect on measurement precision. Yet at least one participant

Table 1. The Percent of Respondents Choosing Each Category

Scenario 1 (NRT)	No Effect	Very Little Effect	Some Effect	Large Effect
What would you expect to be the effect on measurement precision if a **locator** test was used to assign a student to ...				
one level below grade?	9	39	44	9
two levels below grade?	0	17	35	48
more than two levels below?	0	5	23	73
What would you expect to be the effect on measurement precision if a **teacher** assigned a student to ...				
one level below grade?	9	36	36	18
two levels below grade?	0	19	38	43
more than two levels below?	0	5	29	67
Scenario 2 (CRT)				
What would you expect to be the effect on measurement precision if a **teacher** assigned an 8th grade student to...				
the 5th grade test?	0	0	56	44
the 3rd grade test?	0	0	18	82

indicated that linking error would not offset the gain in measurement precision, and several indicated that it would have some effect. The pattern of responses was similar when asked to consider the situation in which a teacher assigned students to test level.

Respondents seemed much more concerned about the impact of the imprecision of the vertical linking with a criterion-referenced test in which scores from a 3rd and 5th grade criterion-referenced test were linked to the 8th grade test through the equipercentile method. For the situation in which an 8th grader takes the 5th grade test, and the score is translated onto the scale of the 8th grade test, the consensus opinion was that there would be some to a large reduction on the precision of performance estimates. For the situation in which an 8th grader takes the 3rd grade test, consensus opinion was that there would be a large reduction in measurement precision.

Validity

The results for the Likert items in the second section are shown in Table 2. As is apparent in the table, the opinions of the experts varied even more on these items than on the reliability items.

For Scenario I, involving a multi-level NRT, 23% of the respondents indicated that combining out-of-level test scores with in-level test scores would have no effect on the interpretation of school-to-school comparisons, whereas all other respondents indicated that it would either somewhat or dramatically reduce the meaning of the scores. When it comes to monitoring

Table 2. Percent of Respondents Selecting Each Category

Scenario 1 (NRT)	Dramatically Enhance	Somewhat Enhance	No Effect	Somewhat Reduce	Dramatically Reduce
What would you expect to be the effect on score interpretation if test scores are used ...					
for school-to-school comparisons?	0	0	23	59	18
to monitor adequate yearly progress of the school?	4	4	18	46	27
to determine whether the student met the passing standard?	5	10	10	19	57
to guide classroom instruction?	24	24	14	29	10
Scenario 2 (CRT)					
for school-to-school comparisons?	0	4	9	46	41
to monitor adequate yearly progress of the school?	4	4	4	41	46
to determine whether the student met the passing standard?	0	10	10	33	48
to guide classroom instruction?	9	14	14	50	14

adequate yearly progress, 8% of the respondents believed that out-of-level testing would improve the meaning of the scores, whereas 73% believed that it would reduce the meaning of the scores. Fifteen percent of the respondents thought that out-of-level testing would enhance the determination of whether a student met the passing standard; 76% thought that the determination of passing would be degraded. Nearly one-half of the participants thought that out-of-level testing would enhance evaluation of classroom instruction, whereas 39% felt that it would degrade the evaluation.

In the second scenario, a CRT developed for 8th graders represented the in-level test, and the out-of-level test was either the 5th grade CRT or the 3rd grade CRT. Scores from the tests were linked through the equipercentile method, and performance was reported on the 8th grade scale. The pattern of responses did not change much for this scenario as for the first. The main difference was the respondents were somewhat less likely to indicate that out-of-level testing would enhance the interpretation of the scores.

Under Sampling in Linking Studies

Scenario III examined the question of whether the disparity between the representation of students with disabilities in out-of-level testing and their representation in vertical linking studies biases the out-of-level testing results for students with disabilities. Most of the students who are tested

BEST COPY AVAILABLE

out of level are students with disabilities, yet they tend to represent only a fraction of the sample in vertical linking studies. When asked what effect this disparity would have on test scores for students with disabilities taking out-of-level tests, participants could identify as many responses as they wanted from four possibilities: "No Effect," "Introduces Bias," "Introduces Measurement Error," and "Other." The percentages of participants choosing each were: No Effect – 9%; Introduces Bias – 46%; Introduces Measurement Error – 58%; Other – 33%.

Narrative Feedback

Three general topics emerged from the respondents' feedback when provided an opportunity to write open-ended comments within the Web-based survey.

Topic 1 – The Difficulty of Setting an Assessment Context

In a previous study that gathered opinions from test and measurement experts about testing students with disabilities out of level (Minnema et al., 2001), participants indicated that they needed an assessment context within which to ground their opinions. Even though the purpose of this survey project was to meet the participants' needs by providing an assessment context within which to respond, they raised concerns about the details in the scenarios. However, there was no overlap in the feedback about scenario content. In other words, each issue raised was raised only one time overall. For instance, one participant commented, *"It would be helpful to know more about the reliability (or conditional standard errors) of each of the [test] forms. If the test is not highly reliable, there is likely to be some regression to the mean effects, particularly if assignment to test forms is correlated with true ability. Different demographic groups may regress toward different means, which will introduce some amount of bias."*

Other respondents suggested ways in which they thought the scenarios could be improved. *"The survey questions ignore [curricular] content. Knowing only about difficulty but not about content, provides too little information to accurately address the survey questions."* Further, *"My problem answering this question is that I reject the scenario. A norm-referenced test cannot be reasonably well aligned [with state standards]. Norm-referenced tests are developed to have broader coverage than would be the case for a standards-based assessment. Moreover, even if it were aligned at one level, it might not be at another level."* Again, no two participants suggested the same improvements to the scenarios.

Topic 2 – Emergence of Points of Agreement and Disagreement

One topic emerged from the narrative comments that points to some convergence of opinion among our participants. Respondents indicated on more than one occasion that *"there is more concern as the number of levels away [from the grade level of enrollment] increases"* when recruiting a sample of students for a linking study. This concern also surfaced when addressing the use of a locator test in assigning levels of an out-of-level test. *"The measurement error increases significantly when the student is assigned to a test two or more levels below [assigned] grade level. This is specifically relevant at the 8th grade or below because of the sharper slope of the learning curve at that stage of educational progress."*

In other instances, respondents qualified their ratings by presenting opposing points of view. For instance, two opinions were raised in comparing the use of a locator test to teachers' judgment in assigning levels of an out-of-level test. Some respondents indicated, *"The research on using locator tests and teacher assignment is about the same. If the test a student is taking doesn't have direct vertical equating, I'm not sure I would trust the equating results."* However, another participant commented, *"I think that we overestimate our ability to place students appropriately and that we consequently underestimate the amount of imprecision that may be introduced. But, a GOOD indicator test is probably more likely to appropriately target [a test level] than teacher judgment."*

Topic 3 — The Role of Student-related Factors and Psychometric Test Characteristics

It is interesting to note that this purposive sample of test and measurement experts considered contextual factors of administering an out-of-level test that could affect the psychometric integrity of the test results. In other words, they did not think about technical issues of a test in a vacuum. For instance, *"The problem is in the appropriateness of the items for the age of the student. If the purpose of dropping a level(s) in testing is to get the best match for the cognitive level of the student, the data may be confounded by the age-context variables (Is a story about a fluffy bunny actually appropriate for a teenaged student?)"* In referring to the validity of the test results when tests are administered at multiple grade levels, another participant wrote, *"The issue is not just of level. For example, a particular neurological condition might impact performance on some items but not on others. So, in some cases the issue is not on the grade level of the content assessed, it is a question of how the content is assessed. Yet, in other situations, it is the interaction of these two factors."* There were also cases where the respondents thought beyond the administration of a statewide assessment to consider such factors as *"misinterpretation of the results by parents and some teachers."* Clearly, the results of this Web-based survey reflect the realities of implementing an out-of-level testing program in educational practice when thinking about the psychometric issues that surround out-of-level testing.

Discussion

This study asked people with expertise in psychometrics and the use of large-scale assessment data to consider possible positive and negative effects of out-of-level testing on test score reliability and validity. Participants were presented several scenarios that were hypothetical examples of realistic large-scale testing programs. The scenarios were brief, but included critical information needed for making judgments. The scenarios varied test type (NRT vs CRT), number of levels below grade level (one, two, or more than two), and the type of linking (vertical scaling within a multilevel NRT programs vs. the equipercentile method in which a sample of students took two levels of the CRT).

The first section of the survey dealt with the issue of measurement precision. Specifically, whether the error introduced in the linking process offset the gain in measurement precision expected when a student is assigned to a test that corresponds to that student's ability. Opinions of experts clearly varied. Generally, the participants indicated that the more levels below grade level a student was tested, the more the linking error offset the precision gained from out-of-level testing.

The *Standards for Educational and Psychological Testing* recommend that test publishers provide detailed technical information on the method by which linking functions were established and on the accuracy of linking functions (APA/AERA /NCME, 1999; Standard 4.11). Test publishers usually do not provide information on linking error, possibly because it is not possible to determine the accuracy of linking functions with real data (Kim & Cohen, 1998). Few studies have examined the amount of error (see Bielinski et al., 2000); additional research is needed.

As important as reliability is and as well understood as it is, it pales in comparison to the importance of valid measurement. The comments of one participant best illustrate the concern:

> *My major concern is not with the reliability of the scores. My major concern is with the validity of the interpretation of the scores, even in a norm-referenced sense. If a state has content standards for grade 8 and the 8th grade test aligns fairly well with that content, how well does the grade 7 or grade 6 test align? Thus, the interpretation of the scores in terms of a student's performance on the content standard may be adequately reliable, but not at all valid.*

In other words, an out-of-level test may not provide an accurate measurement of an individual's standing on the skills that the in-level test was designed to measure. This possibility raises two important concerns: (1) can scores from a lower level be aggregated with scores a higher level without compromising the meaning of the aggregate?; (2) can the score on a lower level test adequately translate into an index of proficiency on the skills measured by the in-level test?

The great variability in the opinions of our participants about the question of whether scores from different test levels can be combined without compromising the meaning of the aggregate scores is consistent with the uncertainty surrounding developmental scales. Most participants indicated that out-of-level testing

Participants also indicated that under-sampling of special populations in linking studies introduces bias and measurement error into the common scale on which test scores are equated. Participants were allowed to select more that one response, and it was evident that many did so, with most indicating that under-sampling introduces bias and measurement error. This finding emerged in the narrative data as well, with one respondent stating:

> *I think publishers and other non-state assessment programs face considerable challenges in recruiting participation in tests, especially norming and other studies, and so must compromise the random-representativeness of their samples a fair amount. . . . Despite these reservations I have to believe that there would be biasing as well as equating/ measurement error effects in the vertical linking.*

This study is a step toward understanding the impact of out-of-level testing on test score reliability and validity. Yet, our findings are constrained by two important considerations. First, our scenarios were not written comprehensively enough to satisfy all of our survey respondents. Instead, participants often qualified their responses by suggesting improvements to the scenarios. It is difficult to build enough detail into a scenario to obtain meaningful evaluation of its effects. Each state has its own unique circumstances, and an amalgam of the issues may not be sufficient to provide responses that can be generalized across locations.

The second constraint is the range of knowledge about the technical psychometric issues surrounding out-of-level testing and scale development. While we employed specific criteria in recruiting participants, it was difficult to ascertain their exact level of understanding and experience. It is likely that some respondents had expertise in other aspects of large scale assessment, and may not have had expertise in the psychometrics, particularly as it pertains to scale development and vertical linking.

The psychometric issues surrounding out-of-level testing, while not resolved with this study were clarified somewhat. The participants indicated that the context is very important and that the number of levels out-of-level is important. Recommendations about out-of-level testing need to consider the purpose of the scores, the type of test, and the type of method used to link scores across tests.

All of these findings point to the difficulty in making recommendations to policymakers and educators who are striving to ensure the best measurement for all students, including students with disabilities. The multiplicity of opinions that emerged in this study underscores the need

for further research and explication regarding the benefits and limitations of out-of-level testing. Perhaps the best way to summarize the current state of opinion regarding the psychometric features of out-of-level testing is to quote one of the participants who stated:

> *Remember this, [out-of-level testing] is not a theoretical, abstract measurement problem that the virtues of scaling can solve... What is possible theoretically does not necessarily mean it will actually work in the real world.*"

References

Bielinski, J., Thurlow, M., Minnema, J., & Scott, J. (2000). *How out-of-level testing affects the psychometric quality of test scores* (Out-of-Level Testing Report 2). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Bourque, L., & Fielder, E. (1995). *How to conduct self-administered and mail surveys.* Thousand Oaks, CA: Sage Publications.

Kim, S-H., & Cohen, A. S. (1988). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22* (2), 131-143.

Minnema, J., Thurlow, M., & Bielinski, J. (2002). *Test and measurement expert opinions: A dialogue about testing students with disabilities out of level in large-scale assessments* (Out-of-Level Testing Report 6). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Minnema, J., Thurlow, M., & Scott, J. (2001). *Testing students out of level in large-scale assessments: What states perceive and believe* (Out-of-Level Testing Report 5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

National Center on Educational Outcomes (NCEO). (2001). *FAQ: Universally designed assessments – NCEO topic area*. Retrieved October 16, 2001, from http://education.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_FAQ.htm

Thurlow, M., Elliott, J., & Ysseldyke, J. (1999). *Out-of-level testing: Pros and cons* (Policy Directions 9). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., & Minnema, J. (2001). *States' out-of-level testing policies* (Out-of-Level Testing Report 4). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendix A

Focus Group Follow-up Survey on Psychometric Impact of Out-of-Level Testing

Focus Group Follow-up Survey on Psychometric Impact of Out-of-level Testing

This survey asks for your expert opinion about how out-of-level testing effects test score reliability and validity. An out-of-level test is defined here as any test taken by a student that is other than the test taken by the majority of students in her/his grade. A context is provided for each question because the conditions within which out-of-level testing occurs may influence your opinions.

Please note: you may change you survey responses at any time before pressing the submit button. However, once you have submitted the survey, all responses will be final and additional changes will not be possible. If you have any questions, or if you run into any problems filling this survey out, please contact John Bielinski at bieli001@umn.edu
Thank you for taking the time to complete this survey.

Test Score Reliability

From item response theory, it can be shown that measurement precision increases as the match between test difficulty and person ability increases. One benefit of out-of-level testing is that test score reliability may be increased for those students who would otherwise earn a very low or a very high score on the in-level (grade level) test. However, when the performance of the students taking an out-of-level test is to be reported on the scale of the in-level test, some form of equating of scales or linking of item parameter estimates is required. Equating may add measurement error to the scaled score. Questions 1-3 ask your opinion on the degree to which the potential gain in measurement precision from out-of-level testing is offset by the measurement error introduced in the equating process.

Scenario I

A state wants to assess math proficiency for all its 8th graders. An off-the-shelf norm-referenced test was chosen that was standardized on an 8th grade sample, and was reasonably well aligned to that state's mathematics standards. The 8th grade test is part of a multi-level testing system in which test scores from any level of the test can be placed onto a common scale. The test publisher conducted vertical scaling/equating studies for this purpose.

Please rate the extent to which the measurement error added in the vertical equating process offsets the gain in measurement precision obtained by giving a student an out-of-level test under each condition.

A.) Assume that a brief locator test was used to assign each student to a test level....

The student was assigned to the test	None	Too little to Warrant concern	Some	A lot
• One level below	A	A	A	A
• Two levels below	A	A	A	A
• More than two levels below	A	A	A	A

B.) Assume that the classroom teacher made the decision as to which level each student should take....

The student was assigned to the test	None	Too little to Warrant concern	Some	A lot
• One level below	A	A	A	A
• Two levels below	A	A	A	A
• More than two levels below	A	A	A	A

Briefly explain your choices

Scenario II

A state wants to assess math proficiency for all its 8th graders. The state has developed a test specifically designed to measure that state's math standards. The state has also developed a math test to assess the math standards for its 5th graders, and one to assess the math standards for its 3rd graders. The state conducted an equating study so that a score from either the 3rd grade or the 5th grade test could be translated into a score on the 8th grade test. A random sample of 500 8th graders took both the 8th grade and the 5th grade test, and another random sample of 500 8th graders took both the 8th grade test and the 3rd grade test. Equipercentile equating was used to translate scores from the 3rd grade and the 5th grade test to the scale of the 8th grade test. A classroom teacher familiar with the student made the decision as to which test to give that student. All scores, regardless of the level of the test a student took were reported on the scale of the 8th grade test.

Please rate the extent to which the measurement error added in the vertical equating process offsets the gain in measurement precision obtained by giving a student an out-of-level test under each condition.

An 8th grader was assigned to the	None	Too little to Warrant concern	Some	A lot
• 5th grade test	A	A	A	A
• 3rd grade test	A	A	A	A

Scenario III

Participants in the vertical equating studies conducted by the large test publishing companies are selected so that they best represent the demographic characteristics and ability levels of the population of students in a particular grade. However, the students who actually take an out-of-level test are likely to differ from their grade level peers in both ability and demographic characteristics. What, if anything, do you consider would be the effect on test results from this dissimilarity between the participants in vertical equating studies and the students who actually take an out-of-level test?

No effect	A
Introduces test score bias	A
Introduces random error	A
Other possible effect	A

BEST COPY AVAILABLE

Test Score Validity

Out-of-level testing may increase measurement precision for low or high scoring students. However, there is concern that it may alter the validity of the test score. Below we are interested in getting your expert opinion about the degree to which an out-of-level test score alters the validity of the data. Because meaningful **evaluation of validity requires** information about how the scores are to be reported and used, we present each set within a specific context. The contexts differ in the way in which test scores are reported and how the results are to be used.

Scenario I

- An off-the-shelf norm-referenced test was used. The test was part of a multi-level testing system in which vertical equating was done so that scores from each test could be placed onto a common scale.
- No student was allowed to take a level of the test more than two levels below the test intended for her/his grade level.
- The percent of students within a grade taking an out-of-level test varied across schools.
- No distinction was made between students getting an out-of-level test and those getting the in-level test.

Select the box that best reflects your opinion as to the effect out-of-level testing (using the context above) has on the validity of the test results. For school level results, assume that scores from out-of-level tests are pooled with those from in-level tests to generate the summary statistic.

Report Format A: At the school level, the percent of students in a grade that met the state standard					
	Effect on Validity				
Scores will be used...	**Dramatically Reduce**	**Somewhat Reduce**	**No Effect**	**Somewhat Enhance**	**Dramatically Enhance**
for school-to-school comparisons.	A	A	A	A	A
to monitor adequately yearly progress. Each school must demonstrate gain in the percent of students meeting the state standard.	A	A	A	A	A

Report Format B: At the school level, the mean scaled score for each grade tested					
	Effect on Validity				
Scores will be used...	**Dramatically Reduce**	**Somewhat Reduce**	**No Effect**	**Somewhat Enhance**	**Dramatically Enhance**
for school-to-school comparisons.	A	A	A	A	A
to monitor adequately yearly progress. Each school must demonstrate a specified gain their mean scaled score.	A	A	A	A	A

Report Format C: Individual student score					
	Effect on Validity				
Scores will be used...	**Dramatically Reduce**	**Somewhat Reduce**	**No Effect**	**Somewhat Enhance**	**Dramatically Enhance**
to determine eligibility for high school graduation. Each student must pass the test to graduate.	A	A	A	A	A
to guide classroom instruction.	A	A	A	A	A

Scenario II

- A state developed test was used. The testing system included a 3rd grade, a 5th grade, and an 8th grade test. A score from a lower level test (e.g. 5th grade test) could be translated into a score on a higher-level test (e.g. 8th grade test). The equating scenario described in the reliability section, scenario II was used.
- An 8th grader could take either the 3rd grade, the 5th grade, or the 8th grade test. A teacher familiar with the student made the determination as to which test was most appropriate.
- The percent of students within a grade taking an out-of-level test varied across schools.
- No distinction was made between students getting an out-of-level test and those getting the in-level test.
- Consider only the 8th grade test results.

Report Format A: At the school level, the percent of students in a grade that met the state standard					
	Effect on Validity				
Scores will be used...	**Dramatically Reduce**	**Somewhat Reduce**	**No Effect**	**Somewhat Enhance**	**Dramatically Enhance**
for school-to-school comparisons.	A	A	A	A	A
to monitor adequately yearly progress. Each school must demonstrate gain in the percent of students meeting the state standard.	A	A	A	A	A

Report Format B: Individual student score					
	Effect on Validity				
Scores will be used...	**Dramatically Reduce**	**Somewhat Reduce**	**No Effect**	**Somewhat Enhance**	**Dramatically Enhance**
to determine eligibility for high school graduation. Each student must pass the test to graduate.	A	A	A	A	A
to guide classroom instruction.	A	A	A	A	A

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA

NCEO is an affiliated center of the Institute on Community Integration

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").